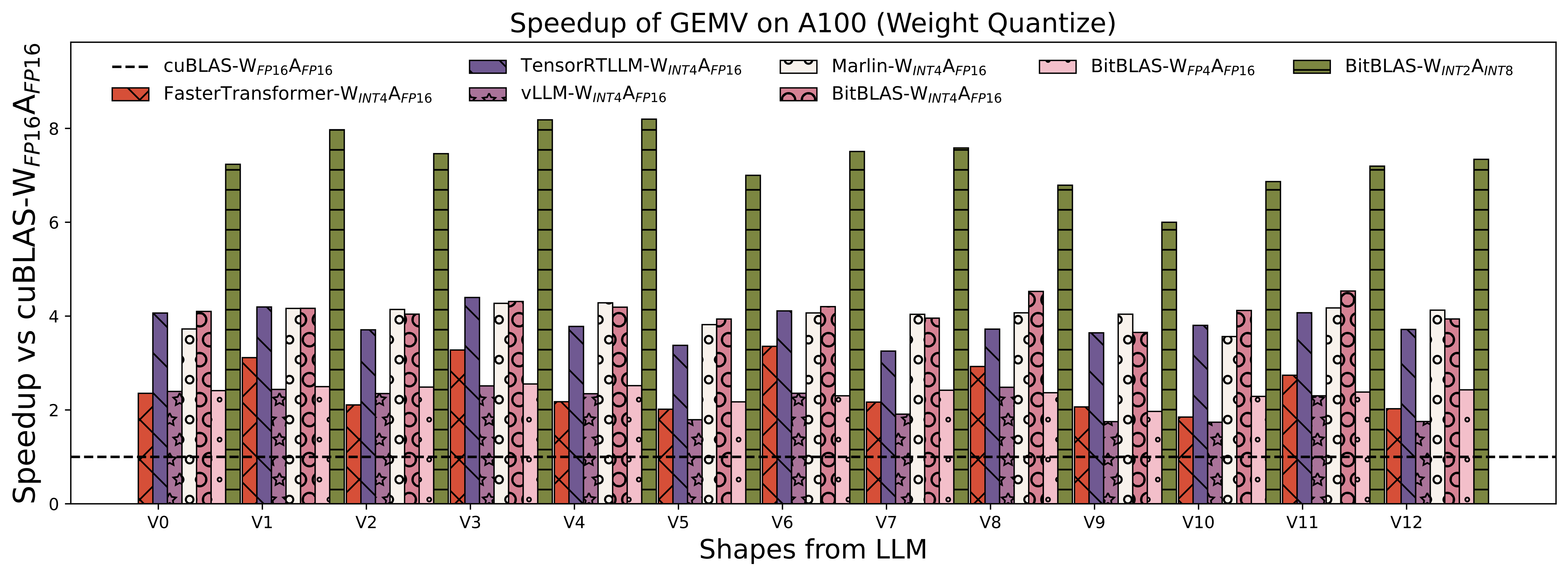
Speedup of GEMV on A100 (Weight Quantize)
cuBLAS-$W_{FP16}A_{FP16}$
FasterTransformer-$W_{INT4}A_{FP16}$
TensorRTLLM-$W_{INT4}A_{FP16}$
vLLM-$W_{INT4}A_{FP16}$
Marlin-$W_{INT4}A_{FP16}$
BitBLAS-$W_{INT4}A_{FP16}$
BitBLAS-$W_{FP4}A_{FP16}$
BitBLAS-$W_{INT2}A_{INT8}$
Speedup vs cuBLAS-$W_{FP16}A_{FP16}$
0
2
4
6
8
V0
V1
V2
V3
V4
V5
V6
V7
V8
V9
V10
V11
V12
Shapes from LLM